ENRON EMAIL ANALYSIS:

BACKGROUND:

This is a Data Engineering Project implemented at a small scale to analyze the email logs generated by Enron Corporation. Enron Corporation was an American energy company that went bankrupt after engaging in massive accounting fraud. As part of federal investigation, government released large number of emails written by Enron's executive team to public. We are using those email logs as source files for this project.

OBJECTIVE:

- Who received the most emails on which day? Please list the top 20 sorted by volume
- Let's label an email as "direct" if there is exactly one recipient and "broadcast" if it has multiple recipients. Identify the top five (5) people who received the largest number of direct emails, and the top five (5) people who sent the largest number of broadcast emails.
- Find the five (5) emails with the fastest response times. A response is defined as a message from one of the recipients to the original sender. The subject line of responses starts with either "RE:" or "FW:" (trimmed whitespace, case insensitive) followed by the original subject line. The response time should be measured as the difference between when the original email was sent and when the response was sent.

APPROACH:

- Data Model:

  Using Dimensional Model on top of RDBMS

- Data Pipelines:

  Using Python and SQL to build the data pipeline.